Practical Name Disambiguation

Kirk Baker, Paula Fearon, Patricia Forcinito, Ian Hutchins and George Santangelo

Office of Portfolio Analysis

Background

The ability to accurately link individuals to their publications, grants, patents, and other works of authorship is critical to performing data-driven analyses of current and emerging areas of research. Efforts to establish unique identifiers for authors are under way (e.g., ORCID, SciENcv), but name disambiguation remains a problem for many analyses, particularly when working with data from multiple sources that is inconsistently formatted. We describe a robust name disambiguation tool that makes minimal assumptions regarding data formatting and degrades gracefully in the absence of reliable data fields.

Methods

Our name disambiguation approach relies on a combination of heuristics and statistical machine learning.

Comprised of three components:

- Citation parsing
- Name parsing
- Name disambiguation

An initial evaluation of the approach using 1.2 million ORCID profiles suggests disambiguation accuracy well over 90%.

Citation Parsing

Citation formats vary across data exported from different sources. Typical author delimiters include comma, semicolon, space, and the word 'and'. Comma may function as either a between- or within-name delimiter. We use heuristics to determine the most likely parse.

Semi-colon delimited author list (ambiguous within-name comma):
Walker, Bailus, Jr.; Kassim, Kunle; Stokes, Lynette Denise

Comma-delimited author list:
Vogan JM, Collins K.

Single-author list (ambiguous within-name comma):
Peek, Richard M., Jr.

Name Parsing

Name formats vary across data exported from different sources with respect to character set, capitalization, the order of family and given names, use of initials vs full names, and within-name delimiter (typically comma or space). We are currently using a maximum entropy-based machine learning model to predictively parse a name into its constituent parts (Family, Given, Initials), irrespective of order. Suffixes are recognized heuristically.

Names are transliterated to basic Latin character set:
Сторожик Марина → {Family:Storozik, Given:Marina}
김 동찬 → {Family: Gim, Given:Dongchan}

Family name followed by initials:
Girosky KE → {Family:Girosky, Initial:K, Initial:E}

All-caps family name followed by given name:
JIANG, YI → {Family:Jiang, Given:Yi}
The model incorporates etymology of Chinese names in order to improve the resolution of given names versus initials.

Given name followed by family name:
Liviu Ungur → {Family:Ungur, Given:Liviu}

Camel case given name followed by family name:
DongHoon Lee → {Family:Lee, Given:Dong, Given:Hoon}

Name Disambiguation

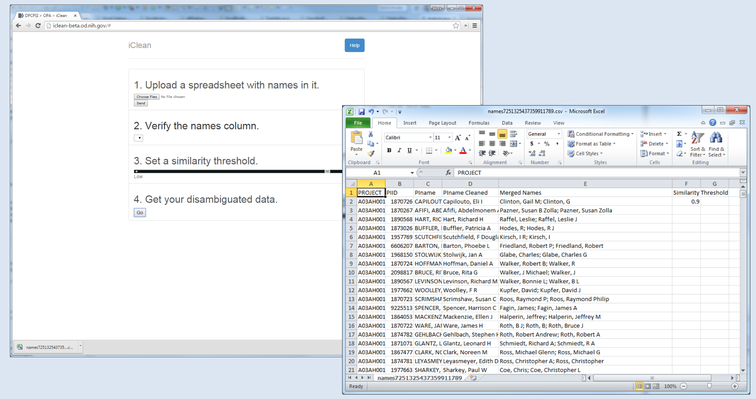
A set of heuristics describes the criteria for matching parsed names and a similarity threshold can be used to further refine results. Common nicknames are recognized and converted to full names (eg, Bill → William). Two names are considered a potential match if:

- Their family names are equal
- The initials of all present given names match in order

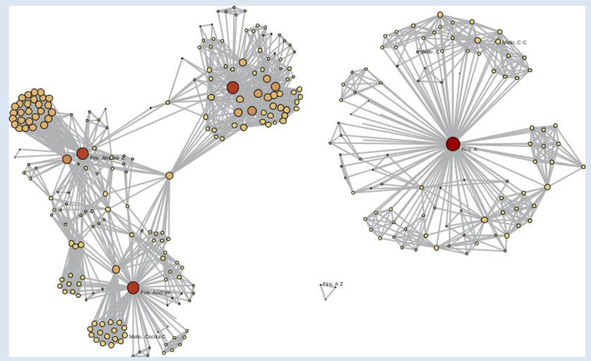
William H. Andrews	Bill Andrews	✓	0.82
Andrews WH	William H Andrews	✓	0.1
Andrews WH	William D Andrews	×	0.1

Current work includes the incorporation of metadata such as institution, coauthors, content similarity, and user feedback into an iterative machine learning model to improve disambiguation quality.

Applications

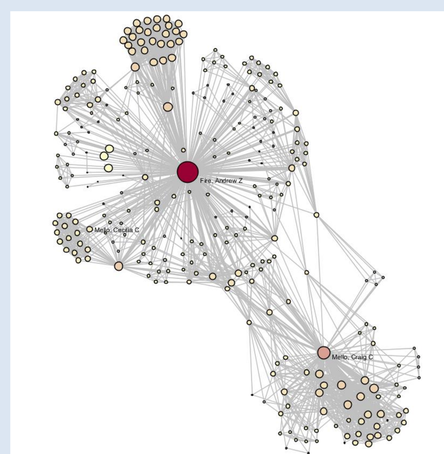


iClean, an OPA web application that performs name disambiguation of user-uploaded spreadsheets.



Coauthor network visualization before name disambiguation.

After name disambiguation, multiple representations of "Andrew Fire" and "Craig Mello" have been merged (along with other authors in the data set), resulting in a more accurate depiction of interactions between researchers.



To be a beta tester of iClean, contact **iClean@mail.nih.gov**